Homework - Day 7

Author: Daniel Ramírez, 2022

Updated by Samuel Hunter, 2024

DESeq2 resources:

https://bioconductor.org/packages/release/bioc/html/DESeq2.html
https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Introduction:

Part 1: FeatureCounts
Practice generating counts using the example bam files on the AWS.

1) Find the BAM files in `/scratch/Shares/public/sread2024/homework_data_files/day7/bam/`
    a. If you want to challenge yourself, try starting from the FASTQ files in `/scratch/Shares/public/sread2024/homework_data_files/day7/fastq` Note that these are single-end samples
    b. Using the hg38 GTF file in /scratch/Shares/public/sread2024/data_files/project/day7/annotations/hg38_ucsc_genes_chr21.gtf, count the reads for all genes in the GTF file.
2) Check the output.
    a. How many samples did you get reads for? 4
    b. How many genes? 351
    c. Check the gene ***UBE2G2***. How many reads did each sample get for this gene? `721,455,468,479`

Part 2: DESeq2

Andrysik et al. ran several experiments which identified a core regulatory program associated with p53 activation across multiple cell lines. In class, we ran differential analysis on RNA-seq data in HCT116 cells. Here, you will run that same pipeline on another cell line, MCF7, from the same paper.

The following files are from the GitHub repository. First update your repo with git pull.

`srworkshop/projectB/day07/homework/featureCounts/MCF7_counts.tsv`
`srworkshop/projectB/day07/homework/featureCounts/MCF7_samples.tsv`

1) Read these files into your R environment. Are these files in the proper format to enter into DESeq2? No, the Length column must be removed

2) Run DESeq2 on these samples, using an experimental design that tests whether the Nutlin-treated samples show any significant differences from the DMSO-treated ones
    a. Generate histograms and boxplots for the normalized counts. Are there any issues with any of the samples? SRR4098435 is not replicating well
    b. Generate a PCA plot, coloring the samples by their treatment group. How do the samples group together? Nutlin samples group together, DMSO samples do not, separated on PC1
3) Generate the DESeq2 statistical results. Use an adjusted p-value cutoff of 0.1.
    a. How many genes were upregulated upon Nutlin treatment? How many were downregulated?
    LFC > 0 (up) : 263, 1.1%
    LFC < 0 (down) : 963, 3.9%
    b. Generate an MA plot and a Volcano plot. Color the significant genes.
    c. What's the top hit in the DESeq2 results? Were there any red flags with this analysis? SRGN- the bad DMSO replicate is probably throwing the analysis off